

Modeling and Testing for Joint Association Using a Genetic Random Field Model*

Zihuai He*, Min Zhang*, Xiaowei Zhan* and Qing Lu**

e-mail: zihuai@umich.edu; mzhangst@umich.edu

**University of Michigan*

***Michigan State University*

Abstract: Substantial progress has been made in identifying single genetic variants predisposing to common complex diseases. Nonetheless, the genetic etiology of human diseases remains largely unknown. Human complex diseases are likely influenced by the joint effect of a large number of genetic variants instead of a single variant. The joint analysis of multiple genetic variants considering linkage disequilibrium (LD) and potential interactions can further enhance the discovery process, leading to the identification of new disease-susceptibility genetic variants. Motivated by the recent development in spatial statistics, we propose a new statistical model based on the random field theory, referred to as a genetic random field model (GenRF), for joint association analysis with the consideration of possible gene-gene interactions and LD. Using a pseudo-likelihood approach, a GenRF test for the joint association of multiple genetic variants is developed, which has the following advantages: 1. considering complex interactions for improved performance; 2. natural dimension reduction; 3. boosting power in the presence of LD; 4. computationally efficient. Simulation studies are conducted under various scenarios. Compared with a commonly adopted kernel machine approach, SKAT, GenRF shows overall comparable performance and better performance in the presence of complex interactions. The method is further illustrated by an application to the Dallas Heart Study.

Keywords and phrases: Joint association; Random field; High-Dimensional test; Complex interaction; Linkage disequilibrium..

1. Introduction

With the advance of high-throughput technologies, high-dimensional genetic data have been widely used in association studies for the identification of genetic variants contributing to common complex diseases. While a large number of genetic variants have been revealed today to be individually associated with complex diseases, they only explain a small proportion of heritability (Manolio, et al., 2009). On one hand, complex diseases are likely influenced by the joint effect of genetic variants through complex biology pathways, given the fact that genes are the functional sets (Wei, et al., 2013). On the other hand, the

*This paper has been submitted for consideration for publication in Biometrics

multiple testing problem occurs when one considers a set of single locus analyses, which dramatically diminishes the power. Therefore, the joint analysis of a functional set of genetic variants simultaneously can further enhance the discovery process, leading to the identification of new genetic variants associated with complex diseases (Chatterjee, et al. 2006). While the conventional linear or logistic regression models can easily be used for joint association analyses, they are subject to several issues, such as multiple-collinearity, when dealing with a large ensemble of dense genetic markers. The exponentially increased number of parameters also making the methods impractical to model two-way or high-order interactions among a large number of genetic variants (Cordell, 2009; Ritchie, et al., 2001).

Several new statistical methods have also been recently developed for joint association analysis. It is worthwhile to note two recently developed methods: the kernel machine based methods (well known as SKAT)(Wu, et al.; 2010; Wu, et al. 2011) and the similarity regression (SIMreg) (Tzeng, et al. 2009). SKAT is developed from a kernel machine or random effect model framework and SIMreg directly models trait similarity as a function of genetic similarity. Both methods significantly reduce the number of regression parameters, making it feasible and computationally efficient to handle high-dimensional variants. In addition, both SKAT and SIMreg use flexible frameworks to exploit and account for linkage disequilibrium (LD) and potential interactions, which further improve the performance of the methods. The two methods have also been extended in several ways. Tzeng, et al. (2011) extended SIMreg to evaluate gene-environment ($G \times E$) interaction and showed the close link between SIMreg and SKAT; Li, et al. (2012) developed another kernel machine based method for gene-gene ($G \times G$) interaction; Maity, et al. (2011) further applied garrote kernel to evaluate a single variant, considering both the main effect and potential interactions with other genetic variants.

In this paper, we propose a novel, random field framework for modeling and testing for the joint association of multiple genetic variants. In this method, we view outcomes as stochastic realizations of a random field on a genetic space and propose to use a random field model, referred to as a genetic random field model (GenRF), to model the joint association. This approach is motivated by development in spatial statistics where outcomes can be viewed as stochastic realizations of a random field on a 2-dimensional space (Cressie, 1993). Thus, our approach can be viewed as a generalization of spatial statistics from a 2-dimensional space to a k -dimensional space. This random field perspective leads to a very distinctive model from the aforementioned regression-based methods including SKAT and SIMreg; specifically, GenRF regresses the response of one subject on responses of all other subjects, which is not common in many fields other than spatial or time-series statistics. Although the random field framework may be unfamiliar to some statisticians, as we demonstrate later, the method can be understood from the intuitive idea that genetic similarity will lead to trait similarity if variants are associated with the trait. Under the GenRF model, testing for the joint association reduces to a test involving a scalar parameter. A testing procedure for the joint association using the pseudo-likelihood method is

developed. The proposed test possesses many appealing features. For example, it is able to exploit LD and interactions between variants to improve power. The method also allows for adjusting weights of rare variants to boosting power. In addition, it is computationally convenient. Our simulation studies and an application to a real data show that the proposed method can achieve comparable or, in the presence of complex interactions, superior performance to SKAT.

The remainder of this article is organized as follows. In Section 2 we set up the notation and introduce related background. We describe the proposed genetic random field model in Section 3.1, the relationship with previous work in Section 3.2, and develop a joint association test under the GenRF model on Section 3.3. The performance of the proposed method is evaluated by simulation studies in Section 4. The proposed method is illustrated by an application to the Dallas Heart Study in Section 5, followed by a discussion in Section 6.

2. Notation and Background

Consider a study where n subjects are sequenced in a region of interest. For subject $i, i = 1, \dots, n$, let \mathbf{G}_i ($p \times 1$) denote the genotype for the p variants within the region and Y_i the trait or phenotype that \mathbf{G}_i is potentially associated to. Additionally, one may also collect other covariates, denoted by \mathbf{X}_i , on each subject including, for example, age, gender, and other demographic and environmental factors. We are interested in studying the joint association between variants \mathbf{G}_i and trait Y_i , possibly adjusted for the effect of \mathbf{X}_i .

For example, if Y_i is continuous, one might model the relationship between variants and the phenotype by a multivariate linear regression model, given by

$$Y_i = \boldsymbol{\alpha}^T \mathbf{X}_i + \boldsymbol{\theta}^T \mathbf{G}_i + \epsilon_i,$$

where \mathbf{X}_i includes an intercept; $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are the coefficients correspondingly. Testing for the joint association of \mathbf{G}_i with Y_i can be achieved by testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$, i.e., $(\theta_1 = \theta_2 = \dots = \theta_p = 0)$. Although simple and intuitive, this approach has several drawbacks; for example, the usual p -degree-of-freedom tests are known to have low power (Goeman, et al., 2006) and it is difficult to include complex, high-dimensional interactions among variants in parametric models. To remedy these issues, SKAT models the effect of \mathbf{G}_i using a semiparametric linear model, i.e.,

$$Y_i = \boldsymbol{\alpha}^T \mathbf{X}_i + h(\mathbf{G}_i) + \epsilon_i, \quad (1)$$

where $h(\cdot)$ is a nonparametric function assumed to lie in a functional space generated by a positive semidefinite kernel function $K(\cdot, \cdot)$; e.g., $K(\mathbf{G}_i, \mathbf{G}_j) = \sum_{k=1}^p \mathbf{G}_{ik} \mathbf{G}_{jk}$ would correspond to a linear model. The form of $h(\cdot)$ is not explicitly specified and instead is implicitly determined by the chosen kernel function. Under this model, testing for the joint association is equivalent to testing $H_0 : h(\mathbf{G}_i) = 0, i = 1, \dots, n$. This is achieved by viewing $[h(\mathbf{G}_1), \dots, h(\mathbf{G}_n)]$ as a random vector with mean zero and covariance $\tau \mathbf{K}$, where \mathbf{K} is an $n \times n$ matrix with the (i, j) -th element equal to $K(\mathbf{G}_i, \mathbf{G}_j)$, and then testing $H_0 : \tau = 0$

by a variance-component score test (Lin, 1997). In summary, SKAT is based on modeling outcomes in a mixed model framework, where the responses Y_i 's are assumed positively correlated (or similar) across i due to the random effect corresponding to $h(\mathbf{G}_i)$.

SKAT improves power partly by allowing for more flexible models in $h(\mathbf{G}_i)$ via the choice of a proper kernel function $K(\cdot, \cdot)$. As explained in Wu, et al. (2011), the kernel function $K(\mathbf{G}_i, \mathbf{G}_j)$ can be interpreted as a measure for genetic similarity in the region of interest between the i -th and j -th subjects and a kernel function better capturing the similarity between individuals and the causal variant effects can increase power. Thus, in the variance-component score test of SKAT, it implicitly assesses the genetic similarity across subjects and also how this genetic similarity leads to positively correlated (or similar) Y_i 's. Instead of modeling trait similarity by a positive association, the SIMreg of Tzeng, et. al. (2009) explicitly defines a measure for trait similarity and directly regresses the trait similarity between each pair of subjects on genetic similarity. SKAT and SIMreg lead to similar test statistics and an explicit connection between SIMreg and variance component score tests was demonstrated by Tzeng, et al. (2009). We will not discuss the SIMreg test hereafter and focus on comparing our method mainly with SKAT.

3. Method

3.1. Genetic Random Field Model

Our method is also motivated by the general idea that, if the genetic variants are jointly associated with a trait, then the genetic similarity across subjects will contribute to the trait similarity. To put it in another way, if variants are jointly associated with the trait, then the response of a subject would be close to the response of other subjects who share similar genetic and possibly other variables. Based on this key idea, we propose to directly model the response of each subject as a function of all other responses and the contribution of other responses to Y_i is weighted by their genetic similarity. This is in contrast to SKAT which models the relationship of Y_i with \mathbf{G}_i for each i as opposed to that of Y_i with all Y_j 's for $j \neq i$.

For simplicity, we temporarily assume Y_i 's are centered (have mean zero) and there are no other adjustment covariates. Specifically, based on the idea discussed above, we model the conditional distribution of Y_i given all other responses as

$$Y_i | \mathbf{Y}_{-i} \sim \gamma \sum_{j \neq i} s(\mathbf{G}_i, \mathbf{G}_j) Y_j + \varepsilon_i, \quad (2)$$

where \mathbf{Y}_{-i} denotes responses for all other subjects except Y_i ; $s(\mathbf{G}_i, \mathbf{G}_j)$ is known weights, weighting the contribution of Y_j on approximating (or predicting) Y_i via their genetic similarity; γ is a non-negative coefficient measuring the magnitude of the overall contribution, further discussed below; and ε_i 's are random errors corresponding to subject i . A proper weight function $s(\mathbf{G}_i, \mathbf{G}_j)$

gives higher value when the two subjects are more similar in terms of their genetic variants and, as discussed below, can be viewed as a measure for proximity of two subjects in a genetic space. The random errors ε_i 's are assumed to be independent and identically distributed with $\text{Normal}(0, \zeta^2)$; robustness of the method to distributions other than normal is discussed in Section 3.3.

A main distinction between model (2) and the usual regression is that (2) models the conditional distribution of Y_i given responses of other subjects, whereas in the usual regression one models the conditional distribution of a subject's response given explanatory variables of the same subject. As the usual regression is useful for predicting or approximating the response of a subject using his/her covariates, model (2) is useful for approximating the response of a subject using responses of other subjects, where the similarity in responses is due to similarity in genetic variants if variants are associated with the response. The coefficient γ indicates the magnitude of the trait similarity as a result of genetic similarity. Thus, γ can also be interpreted as a measure for the magnitude of the joint association of \mathbf{G}_i with Y_i . Specifically, if \mathbf{G}_i is not associated with Y_i , then regardless of how similar subject i is to other subjects in terms of their genetic variants, the response Y_i is independent of all other Y_j 's for $j \neq i$; that is, $\gamma = 0$. On the contrary, if \mathbf{G}_i is strongly associated with Y_i , then one may expect the response of Y_i can largely be predicted by responses of subjects having the same or similar genetic variants and a large γ indicates a strong joint association. Therefore, we can test the joint association of genetic variants with the trait by testing a null hypothesis involving a single parameter, i.e., $H_0 : \gamma = 0$.

Models like (2), where responses are regressed on responses themselves, are referred to as auto-regressive models and, although less commonly used in genetic and biomedical studies, are commonly used in spatial statistics and in modeling time-series. In this article, we propose to view the response as a random field on a genetic space, and from this fresh perspective, model (2) is formally a conditional auto-regressive (CAR) model (Cressie, 1993).

A random field is a generalization of the notation of a stochastic process (Adler and Taylor, 2007). Informally, a stochastic process is a set of random variables indexed by integers or real numbers; for example, a continuous time-series W_t , $t \in T$, is a stochastic process with an index set $T = \mathcal{R}$. A random field can be defined in more general spaces with the index set being an Euclidean space of dimension greater than one or other spaces. For example, in spatial statistics, crop yields of regions can be viewed as a realization of a random field defined in a two-dimensional space, denoted by $W_{s,t}$ where s and t indicate the (latitudinal and longitudinal) location of a region. Regions that are closer in location have more similar crop yields if spatial correlation exists. Specifically, for our problem, we may view observed responses as realizations of a random field defined in a p -dimensional space of the p genetic variants; that is, corresponding to each "location" in the p -dimensional genetic space (equivalently each vector value that \mathbf{G}_i may take), there is a random response variable associated with it, denoted by $Y_{G_{i1}, \dots, G_{ip}}$ in a slight abuse of notation. Similarly, responses from locations that are "closer" in the genetic space are expected to be more similar

if the genetic association exists. In this sense, our model is a generalization of the auto-regressive model in time series analysis (one dimensional) and spatial statistics (two dimensional). Models like (2) were firstly studied in the seminar work of Besag (1974) for random fields and we will term our model (2) as a genetic random field (GenRF) model. As a matter of fact, the GenRF model is closely related to the conditional auto-regressive model in spatial statistics (Cressie, 1993); that is $s(\mathbf{G}_i, \mathbf{G}_j)$ analogously defines the proximity of neighbor \mathbf{G}_j to \mathbf{G}_i and γ is the counterpart of a spatial dependence parameter. However, we note that the usual tests of spatial dependence, for example, the Cliff-Ord-test (Cliff and Ord, 1972) and the Lagrange Multiplier test (Burrige, 1980), do not apply in our setting to test for the joint association of variants, as discussed in Section 6.

We have yet to define a measure for “closeness” in the genetic space. Suppose each component of \mathbf{G}_i records the number of minor alleles in a single locus and takes on values $\{0, 1, 2\}$, respectively, corresponding to three possibilities $\{AA, Aa, aa\}$. Then a sensible measure for closeness or similarity is the so called identity-by-state (IBS) (Wu, et al., 2010), defined as

$$s(\mathbf{G}_i, \mathbf{G}_j) = \sum_{k=1}^p \{2 - |G_{ik} - G_{jk}|\}.$$

That is, the IBS measures the number of alleles in the region of interest shared by two individuals; for example, in the single locus case ($p = 1$), $s(AA, AA) = 2$, $s(Aa, aa) = 1$, $s(AA, aa) = 0$. The overall similarity between two loci sequences are the sum of shared alleles in all loci in the region of interest between two subjects. Other measures for closeness in the genetic space rather than IBS are also possible, e.g., the other kernel functions discussed in Wu, et al. (2011), providing flexibility in our GenRF model. Similar to SKAT, our GenRF model can also incorporate weights to increase the importance of rare variants. Specifically, one can define $s(\mathbf{G}_i, \mathbf{G}_j) = \sum_{k=1}^p w_k \{2 - |G_{ik} - G_{jk}|\}$, where w_k is a prespecified weight for variant k ; see Wu, et al., (2011) for more discussions on w_k .

The above discussion has focused on the situation where no covariate adjustment is required. If adjustment for other factors, for example, environmental factors, is needed, a natural extension of model (2) is given by

$$Y_i | \mathbf{Y}_{-i}, \mathbf{X}_i \sim \beta^T \mathbf{X}_i + \gamma \sum_{j \neq i} s(\mathbf{G}_i, \mathbf{G}_j) (Y_j - \beta^T \mathbf{X}_j) + \varepsilon_i. \quad (3)$$

An intercept term is included in \mathbf{X}_i and, as a result, in (3) Y_i 's are not required to be centered. Under this model, testing for the joint association of \mathbf{G}_i with Y_i after adjusting for other factors is also equivalent to testing $H_0 : \gamma = 0$. We will mainly focus on this more general form of the GenRF model in the development of a testing procedure. For simplicity, the matrix form of GenRF model is given by

$$\mathbf{Y} | \mathbf{Y}_{-}, \mathbf{X} = \mathbf{X}\beta + \gamma \mathbf{S}(\mathbf{Y} - \mathbf{X}\beta) + \boldsymbol{\varepsilon}, \quad (4)$$

where \mathbf{Y} is $(Y_1, \dots, Y_n)^T$; \mathbf{X} is an $n \times q$ matrix defined as $(\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T$; $\boldsymbol{\varepsilon} \sim \text{Normal}(0, \zeta^2 \mathbf{I}_{n \times n})$; and \mathbf{S} is an $n \times n$ symmetric matrix with zeros on the diagonal and the (i, j) -th element $s(\mathbf{G}_i, \mathbf{G}_j)$ for $i \neq j$, as only the pairs of $\{i \neq j\}$ are involved in the model.

3.2. Relationship with SKAT

We further compare the GenRF model with SKAT. As we have commented previously, SKAT can also be understood in a similar fashion, i.e., if \mathbf{G}_i is jointly associated with Y_i , then subjects having similar genetic variants have similar positively correlated responses. In SKAT, the similarity in responses is essentially modeled in a mixed model framework where a random effect, say, $h(\mathbf{G}_i)$, induces positive correlation among responses. Also in SKAT the similarity in genetic variants (or equivalently the kernel function) does not appear explicitly in the assumed model (1) but is implicitly related to $h(\mathbf{G}_i)$ according to the kernel machine regression theory. In contrast, our model in (2) models the similarity in responses via conditional expectations other than correlations and the genetic similarity is incorporated in the model explicitly.

It is easy to see that, in SKAT when $h(\mathbf{G}_i), i = 1, \dots, n$, are treated as following a multivariate normal distribution, model (1) leads to

$$\mathbf{Y}|\mathbf{X} \sim \mathbf{X}\boldsymbol{\alpha} + \mathbf{u}, \quad \mathbf{u} \sim N(0, \sigma^2 \mathbf{I} + \tau^2 \mathbf{K}),$$

where \mathbf{u} is an n -dimensional random column vector; and \mathbf{I} is an $n \times n$ identity matrix. In contrast, according to the factorization theorem of Besag (1974), our GenRF model in (3) leads to the following joint distribution, i.e.,

$$\mathbf{Y}|\mathbf{X} \sim \mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \quad \mathbf{v} \sim N(0, \zeta^2(\mathbf{I} - \gamma \mathbf{S})^{-1}), \quad (5)$$

where \mathbf{v} is an n -dimensional random column vector. Note, the coefficient γ used for describing the conditional expectation of Y_i given others in model (2) actually describes the correlations among Y_i 's. It is clear that, under the null hypothesis that there is no association between \mathbf{G}_i and Y_i , i.e., $\tau = 0$ in SKAT or $\gamma = 0$ in GenRF, the two models are equivalent and Y_i 's are uncorrelated as the covariance matrices are diagonal. Moreover, if $\tau > 0$ or $\gamma > 0$ in the corresponding model, both SKAT and GenRF state that Y_i 's are positively correlated as a result of having similar genetic variants associated with the trait. However, the two models provide different parameterizations of the covariance matrix and consequently the two methods model the magnitude of the joint association differently, albeit both via a scalar parameter. Moreover, though one can adopt the same IBS similarity for both SKAT and GenRF, \mathbf{K} and \mathbf{S} are still different on the diagonal, where the diagonal elements of \mathbf{S} are zeros. The two parameterizations may result in different sensitivity in detecting departures from the null hypothesis. Therefore, one might expect the two models lead to testing procedures with different efficiency in testing the genetic effect. Next we develop a testing procedure based on the proposed GenRF model and the proposed test is compared with SKAT by simulation studies in Section 4.

3.3. Genetic Random Field Test

In this subsection, we focus on developing a test for the null hypothesis $H_0 : \gamma = 0$ based on model (3), referred to as the genetic random field test. Model (3) states that, given responses from all other subjects and covariates \mathbf{X}_i , the conditional distribution of Y_i is normal with mean $\beta^T \mathbf{X}_i + \gamma \sum_{j \neq i} s(\mathbf{G}_i, \mathbf{G}_j)(Y_j - \beta^T \mathbf{X}_j)$ and variance ζ^2 . We construct the pseudo-likelihood according to Besag (1975) as

$$L_{pd} = \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi\zeta^2}} \exp \left[-\frac{1}{2\zeta^2} \left\{ Y_i - \beta^T \mathbf{X}_i - \gamma \sum_{j \neq i} s(\mathbf{G}_i, \mathbf{G}_j)(Y_j - \beta^T \mathbf{X}_j) \right\}^2 \right] \right\},$$

which is a product of the conditional densities of Y_i across i . Also according to Besag (1975), assuming β is known, one may estimate γ by the maximum pseudo-likelihood method. It is easy to see that the maximum pseudo-likelihood estimator for γ can be obtained by minimizing $\sum_{i=1}^n \{Y_i - \beta^T \mathbf{X}_i - \gamma \sum_{j \neq i} s(\mathbf{G}_i, \mathbf{G}_j)(Y_j - \beta^T \mathbf{X}_j)\}^2$, which in matrix notation is equal to

$$\{(I - \gamma \mathbf{S})(\mathbf{Y} - \mathbf{X}\beta)\}^T (I - \gamma \mathbf{S})(\mathbf{Y} - \mathbf{X}\beta).$$

The minimization leads to an estimator for γ given by

$$\Rightarrow \tilde{\gamma} = \frac{(\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{S}(\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{S}^2(\mathbf{Y} - \mathbf{X}\beta)}. \quad (6)$$

Intuitively one expects that a large value of $\hat{\gamma}$ would give us evidence to reject the null hypothesis that $\gamma = 0$. In practice, β is unknown. We propose to replace β by its least square estimator $\hat{\beta}$ under the null hypothesis $H_0 : \gamma = 0$, i.e., $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Substitute $\hat{\beta}$ into the expression for $\tilde{\gamma}$ and straightforward algebra leads to the final test statistic:

$$\hat{\gamma} = \frac{\mathbf{Y}^T \mathbf{B} \mathbf{S} \mathbf{B} \mathbf{Y}}{\mathbf{Y}^T \mathbf{B} \mathbf{S}^2 \mathbf{B} \mathbf{Y}}, \quad (7)$$

where $\mathbf{B} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Again a large value of $\hat{\gamma}$ would support the rejection of the null hypothesis.

We next show how the p-value for testing $\gamma = 0$ can be obtained based on the test statistic $\hat{\gamma}$; i.e., we would like to calculate the probability of $\hat{\gamma}$ greater than the observed value of the statistic under the null hypothesis. As \mathbf{S} is not a diagonal or block-diagonal matrix, regular asymptotic argument does not apply. Alternatively, we propose the following procedure to find the p-value by using the exact mixture Chi-square distribution.

Suppose η is the observed value of the test statistic $\hat{\gamma}$. Since $\mathbf{B} \mathbf{S}^2 \mathbf{B}$ is positive-definite, we have

$$P_{H_0} \left(\frac{\mathbf{Y}^T \mathbf{B} \mathbf{S} \mathbf{B} \mathbf{Y}}{\mathbf{Y}^T \mathbf{B} \mathbf{S}^2 \mathbf{B} \mathbf{Y}} > \eta \right) = P_{H_0} \left((\mathbf{B} \mathbf{Y})^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{B} \mathbf{Y} > 0 \right)$$

As it is assumed that $\varepsilon_i \sim N(0, \zeta^2)$, i.i.d. across i , it follows that $\mathbf{BY} \sim N(0, \zeta^2 \mathbf{B}^2)$ under the null hypothesis. On the other hand, the statistic $\hat{\gamma}$ in (7) is ancillary to ζ^2 because ζ^2 in the numerator and denominator will cancel out. Therefore, the above equation becomes

$$P_{H_0} \left((\mathbf{BY})^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{BY} > 0 \right) = P \left(\mathbf{Z}^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{Z} > 0 \right),$$

where \mathbf{Z} is an $n \times 1$ random vector following $N(0, \mathbf{B}^2)$. Applying standard results on the distribution of quadratic form of normal random variables, we have

$$\mathbf{Z}^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{Z} \sim \sum_i^n \lambda_i \Phi_i^2,$$

where Φ_i 's are i.i.d random variables with χ_1^2 distribution, and $\{\lambda_i\}$ are the eigenvalues of $\mathbf{B}(\mathbf{S} - \eta \mathbf{S}^2) \mathbf{B}$. The final p-value can be obtained by Davies' exact method (1980) for the weighted summation of independent Chi-square variables, similar to the p-value calculation used in SKAT (Wu, et al., 2011).

The proposed test has several appealing properties. First, due to the analytical form of the test statistic, the computational burden is well controlled. Second, as $\hat{\gamma}$ in (7) is ancillary to ζ^2 , there is no need to plug in a consistent estimator for ζ^2 as SKAT did. Third, similar to SKAT, the proposed method improves power by exploiting linkage disequilibrium and allowing for possible complex interactions among variants. Linkage disequilibrium can cause correlations between variants, especially when we consider nearby loci. Considering similarity in variants can naturally reduce the degree of freedom. In the extreme case where components of \mathbf{G}_i are "perfectly correlated", the similarity argument will consider the whole set as a single variable, whereas the typical linear regression will have p -degree of freedom. In addition, genetic variants involved in the disease pathway are more likely to interact with each other than contribute to risk individually, known as the epistatic variants effect. Specifying two-way interactions in a set of loci is a challenging high-dimensional problem and the situation gets even worse in modeling higher order interactions. Since in our GenRF model we do not directly model the relationship of \mathbf{G}_i with Y_i , the difficulty of modeling complex interactions are circumvented and the interaction effect is naturally included through measuring genetic similarity. Finally, as SKAT, the proposed GenRF test can boost power of testing rare variants by increasing their weights by specifying w_k appropriately for variant k .

The derivation of the GenRF test given above is built on the normal distribution assumption. Asymptotically, the proposed test is robust to distributions other than normal. Consider $P_{H_0} \left((\mathbf{BY})^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{BY} > 0 \right)$, where it is now assumed \mathbf{Y} follows an arbitrary distribution. The random quantity $(\mathbf{BY})^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{BY}$ is a quadratic form of \mathbf{BY} (with mean 0) with matrix $\mathbf{A} = (\mathbf{S} - \eta \mathbf{S}^2)$. Rotar (1973) proved that under sufficiently weak conditions on matrix \mathbf{A} and for large n , $P_{H_0} \left((\mathbf{BY})^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{BY} > 0 \right)$ is close to $P_{H_0} \left(\mathbf{Z}^T (\mathbf{S} - \eta \mathbf{S}^2) \mathbf{Z} > 0 \right)$, where \mathbf{Z} follows $N(0, \mathbf{B}^2)$ as defined before. In addition, Gotze and Tikhomirov

(1999) gave an upper bound on $\sup_x |P_{H_0}((\mathbf{B}\mathbf{Y})^T \mathbf{A}\mathbf{B}\mathbf{Y} < x) - P_{H_0}(\mathbf{Z}^T \mathbf{A}\mathbf{Z} < x)|$. These properties lead to the natural robustness of the GenRF test as long as $\mathbf{B}\mathbf{Y}$ has expectation zero under the null hypothesis, which is true since the least squares estimator $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ is unbiased for the mean of \mathbf{Y} under the GenRF model when $\gamma = 0$. Our simulation studies in Section 4 further illustrate this robustness. We comment that, as the score test in SKAT is of similar quadratic form, one would expect that SKAT may share this property as well.

4. Simulation

Simulation studies under various scenarios are conducted to evaluate the performance of GenRF test and to compare it with SKAT. In both methods, we adopt the IBS kernel in the corresponding matrix \mathbf{S} or \mathbf{K} and set $w_k = 1$ for each k . Therefore, results from the two methods are comparable.

Three sets of simulations are conducted to evaluate the performance of GenRF test 1) under different levels of LD, 2) under different levels of interaction effect, and 3) under different distributions of the response variable.

In the first set of simulations, data are simulated under different levels of LD effect. For each Monte Carlo data set, genotypes for $p = 10$ loci are simulated for $n = 100$ subjects. To simulate the LD effect, the haplotype is simulated one by one for each locus with the minor allele frequency 0.3 and the haplotypes of each adjacent pair of alleles are correlated with a correlation coefficient ρ with $\rho = 0, 0.1, 0.2, \dots, 0.9$ respectively for each scenario. Genotypes are then generated by summing up two haplotype vectors. This way, all the loci are positively correlated with others in the loci set. The response variable is generated according to the following model

$$Y_i = aG_{i5} + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, 1),$$

$a = 0$ or 0.5 . That is, when $a = 0.5$, the 5th variant is associated with the trait.

In the second set of simulations, data are generated such that complex interaction effect exists. For each Monte Carlo data set, we take $p = 10$, $n = 100$, the minor allele frequency=0.3, and the LD parameter $\rho = 0.4$. Two distributions are considered. In the normal distribution case, the response is related to variants according to the following model,

$$Y_i | \mathbf{G}_i = b \sum_{k=1}^9 G_{i1} G_{ik+1} + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, 1), \quad (8)$$

and b is set to 0, 0.01, 0.02, \dots , 0.1, respectively, in each scenario. In the exponential distribution case, responses are generated as exponential random variables with rate $\lambda = b \sum_{k=1}^9 G_{i1} G_{ik+1}$ with b equal to 0, 0.02, 0.04, \dots , 0.2, respectively. We see that these models contain only interactions but not main effect of each locus.

In the third set of simulations, we further evaluate the robustness of the GenRF test to distributions other than normal. The setup is the same as that

in the first set of simulations with $\rho = 0.4$ except that Y_i is generated according to generalized linear model with a linear predictor aG_{i5} and the canonical link function from the following distributions: Standard Normal, Exponential, Binary. The coefficient a is set to 0.6, 1.1 and 2.5 respectively for each of the distributions. For Mixture Normal, we generate two normal distributions with mean difference 10 with equal proportions. The coefficient for Mixture Normal is 2.7.

For each simulated data set, we test the joint association of variants using three methods: the proposed GenRF test, SKAT and the usual method based on a linear regression model including only main effects of variants. We note that data are not generated according to GenRF models and, therefore, these simulations should not particularly favor the proposed GenRF test, allowing for a fair comparison among methods.

Table 1 shows results for the first set of simulations with LD effect ranging from low to high. All the three tests achieve the type I error rate close to the nominal level. When LD effect does not exist or is low, e.g., $\rho < 0.5$, the test based on a linear regression model is most powerful as expected. However, when the LD effect is moderate or high, both the GenRF test and SKAT have higher or even substantially higher power than the linear regression based test by borrowing information from other loci. The power of the GenRF test is comparable to or slightly higher than that of SKAT.

Table 2 shows results when there are complex interactions between variants but no main effects. In these scenarios, the linear regression method has low power in detecting the joint association. Both GenRF test and SKAT have much larger power. Moreover, the proposed GenRF test has significantly larger power than SKAT in detecting the joint association effect when complex interactions exist, at least in the scenarios considered here.

Table 3 shows the robustness of the proposed GenRF test to distributions other than normal. GenRF test achieves the type I error rate close to the nominal level even when the distribution of the response is not normal; the same holds for SKAT. Similar to results in Table 1, the performance of GenRF test is comparable to that of SKAT when there is a single risk locus.

5. Application

We applied our method to the Dallas Heart Study (Browning et al., 2004; Victor et al., 2004), which was used previously for illustration of association tests (Liu and Leal, 2010; Wu, et al., 2011). The Dallas Heart Study is a population-based, multi-ethnic study where 3551 residents are recruited. For each subject, Lipids and glucose metabolism have been measured. Individuals who have diabetes mellitus, alcohol dependency or have taken lipids lowering drugs are excluded as these factors may confound the interpretation of associations. In this re-sequencing study, 348 sequence variations in the coding regions of the four genes, ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6 are discovered. Most of these variants (86%) are rare with the minor allele frequency less than 1%.

TABLE 1

Simulation results under different levels of LD effects. Parameter ρ indicates the correlation coefficient of two adjacent loci. GenRF: the genetic random field test; SKAT: the sequential kernel association test of Wu, et al. (2011); Linear: the F-test based on a linear regression regression including only main effects. Power: rejection rate when the genetic association exists; Type I: type I error rate, i.e., rejection rate when data are generated under the null model.

Method		Different Levels of LD effect (ρ)									
		No LD	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GenRF	Power	0.435	0.446	0.450	0.508	0.487	0.556	0.634	0.676	0.724	0.804
	Type I	0.053	0.042	0.047	0.048	0.055	0.050	0.045	0.040	0.049	0.061
SKAT	Power	0.420	0.446	0.434	0.496	0.476	0.552	0.616	0.660	0.711	0.800
	Type I	0.050	0.046	0.048	0.041	0.045	0.046	0.048	0.052	0.046	0.062
Linear	Power	0.500	0.516	0.500	0.526	0.504	0.496	0.514	0.493	0.500	0.518
	Type I	0.056	0.042	0.054	0.050	0.058	0.050	0.043	0.048	0.049	0.060

We assessed the association between ANGPTL gene families and two traits, specifically high-density lipoprotein (HDL) and triglyceride, using the proposed GenRF test and SKAT, both with the IBS kernel. Except for the association testing procedure, our analysis is otherwise similar to the original approach (Romeo, et al., 2007) that has discovered the association between ANGPTL4 gene and the level of HDL and triglyceride. For testing the association between ANGPTL4 and HDL, both GenRF test and SKAT showed comparable and marginal evidence for the association between ANGPTL4 and HDL (p-value: 0.085 and 0.060 respectively); however, the p-values are not significant at the level of 0.05. Both methods gave strong evidence for the ANGPTL4 and triglyceride association and the evidence from SKAT is stronger for this particular association (p-values: 0.011 and 1.65×10^{-3}). ANGPTL5 may also be potentially associated with triglyceride (Liu and Leal, 2012; Romeo, et al., 2009). In our analysis, our GenRF test provided marginal evidence to support this association (p-value: 0.071) while SKAT did not (p-value: 0.353). More results are shown in Table 4. Overall, in this application, the proposed GenRF test and SKAT have comparable performance.

6. Discussion

In this article, we have proposed a novel framework for modeling and testing for the joint association of genetic variants with a trait from the perspective of viewing the response as a random field on a genetic space. A random field generalizes the concept of a stochastic process and random field models have widespread applications in areas such as imaging analysis, spatial statistics and

TABLE 2
Simulation results under different levels of interaction effects where the response follows normal distributions or exponential distributions. b is coefficient of the interaction effect explained in Section 4. Other entries as in Table 1.

Method		Different Levels of Interaction Effects (b)									
		Normal Distribution									
		$b = 0.01$	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
GenRF	Power	0.063	0.120	0.176	0.267	0.395	0.539	0.706	0.816	0.853	0.948
	Type I	0.050	0.054	0.053	0.042	0.052	0.049	0.056	0.054	0.046	0.050
SKAT	Power	0.054	0.088	0.122	0.177	0.282	0.392	0.555	0.673	0.737	0.884
	Type I	0.043	0.048	0.044	0.046	0.048	0.044	0.052	0.050	0.041	0.052
Linear	Power	0.050	0.068	0.092	0.132	0.183	0.270	0.391	0.488	0.570	0.764
	Type I	0.047	0.054	0.053	0.042	0.052	0.048	0.050	0.052	0.050	0.056
		Exponential Distribution									
		$b = 0.02$	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
GenRF	Power	0.075	0.125	0.212	0.353	0.486	0.557	0.636	0.740	0.784	0.808
	Type I	0.046	0.053	0.050	0.058	0.052	0.052	0.056	0.040	0.048	0.045
SKAT	Power	0.056	0.090	0.155	0.238	0.340	0.416	0.484	0.600	0.649	0.678
	Type I	0.044	0.052	0.047	0.048	0.048	0.054	0.048	0.045	0.052	0.044
Linear	Power	0.058	0.062	0.088	0.127	0.182	0.218	0.277	0.364	0.391	0.410
	Type I	0.051	0.052	0.052	0.057	0.054	0.048	0.054	0.046	0.050	0.048

TABLE 3
Simulation results under different distributions of the response variable. * indicates results are unavailable due to “sample size is small, need small sample adjustment.” Other entries as in Table 1.

Method		Distribution			
		Normal	Exponential	Mixture Normal	Binary
GenRF	Power	0.725	0.636	0.582	0.646
	Type I	0.051	0.052	0.056	0.046
SKAT	Power	0.726	0.655	0.582	*
	Type I	0.047	0.046	0.046	*
Linear	Power	0.728	0.572	0.568	0.559
	Type I	0.054	0.056	0.054	0.050

TABLE 4
Application to Dallas Heart Study. * indicates p -value is less than $\alpha = 0.05$.

Method	P-value			
	HDL			
	ANGPTL3	ANGPTL4	ANGPTL5	ANGPTL6
GenRF	0.8129	0.0851	0.0009*	0.6803
SKAT	0.6395	0.0596	0.0212*	0.3313
	Triglyceride			
	ANGPTL3	ANGPTL4	ANGPTL5	ANGPTL6
GenRF	0.0136*	0.0105*	0.0716	0.3024
SKAT	0.0020*	0.0017*	0.3527	0.6450

so on. Specifically, in our genetic random field model, we view that there is a random response variable associated with each value of the p -dimensional vector that corresponds to the p variants. An analogy with spatial statistic is helpful to view this vividly, i.e., each vector value of the p variants can be analogously viewed as a “location” in a space in spatial statistics and we term it the p -dimensional genetic space. Our GenRF model is closely related to the conditional autoregressive model in spatial statistics and the parameter γ for describing the association of variants with a trait is the counterpart of a spatial correlation parameter for quantifying spatial autocorrelation. However, as we have mentioned earlier, regular tests for spatial correlation cannot be applied in our setting to test for the genetic association. The reason is that the matrix \mathbf{S} in our GenRF model does not satisfy the regularity condition usually assumed in spatial statistics for deriving the asymptotic distribution, as each subject has infinite neighbors in the genetic space and \mathbf{S} is a dense matrix.

Although motivated from very different perspectives, the proposed GenRF model shares similar features with SKAT. Both methods are based, explicitly or implicitly, on the idea that, if the variants of interest are associated with a trait, then subjects with similar variants have similar (or positively correlated) responses. The proposed GenRF method models the similarity in responses by an autoregressive model for a random field, whereas in SKAT similarity in responses are described by positive correlations induced by random effects in the framework of a mixed effect model. As discussed in Section 3.2, actually both methods lead to correlated responses under the alternative hypothesis that variants are associated with the trait but provide different parameterizations of the covariance. Based on the GenRF model, a test for genetic associations is developed and this test shares many of the appealing features of SKAT. The proposed GenRF test is based on testing a null hypothesis involving a single scalar pa-

parameter, allowing it to exploit LD to improve power. When the LD effect is moderate or high, our simulations show that the GenRF test achieves much higher power than the method based on a linear regression. Similar to SKAT, the GenRF model is flexible enough to allow for complex interaction effects. Our simulations demonstrate that the GenRF test is even much more powerful than SKAT in the presence of complex interaction effects. Moreover, as SKAT, pre-specified variant-specific weights can be incorporated into the model and test to boost power for rare variants. Finally, the GenRF test is computationally easy to implement. In summary, the GenRF test is an appealing alternative to SKAT for testing the joint association of variants with a trait. Based on our simulations, it can achieve overall comparable performance to and sometimes even much better performance than SKAT.

Acknowledgements

The authors would like to thank Jonathan Cohen for the permission to use the Dallas Heart Study data and Dajiang Liu for preparing the data. The authors also highly appreciate Michael Boehnke's comprehensive suggestions, and thank for the valuable comments from Xihong Lin, Seunggeun Lee, William Wen, Veronica Berrocal, Laura Scott, Lu Wang, Dajiang Liu, Bhramar Mukherjee and Hui Jiang.

References

- [1] Adler, R. J. and Taylor, J. E. (2007). Random Fields and Geometry. Springer, New York.
- [2] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B* **36(2)**, 192–236.
- [3] Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *The Statistician* **24(3)**, 179–195.
- [4] Browning, J. D., Szczepaniak, L. S., Dobbins, R., Nuremberg, P., Horton, J. D., Cohen, J. C., Grundy, S. M., and Hobbs, H. H. (2004). Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. *Hepatology* **40(6)**, 1387–1395.
- [5] Burridge, P. (1980). On the Cliff-Ord Test for Spatial Correlation. *Journal of the Royal Statistical Society. Series B* **42(1)**, 107–108.
- [6] Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *The American Journal of Human Genetics* **79(6)**, 1002–1016.
- [7] Cliff, A. and Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis* **4**, 267–284.
- [8] Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics* **6**, 392–404.
- [9] Cressie, N. (1993). Statistics for spatial data. Wiley, New York.

- [10] Davies, R. (1980). The distribution of a linear combination of chi-square random variables. *Applied Statistics* **29**, 323–333.
- [11] Goeman, J. J., Van De Geer, S. A., and Van Houwelingen, H. C. (2006). Testing against a High Dimensional Alternative. *Journal of the Royal Statistical Society. Series B* **68(3)**, 477–493.
- [12] Gotze, F. and Tikhomirov, A. N. (1999). Asymptotic distribution of quadratic forms. *The Annals of Probability* **27(2)**, 1072–1098.
- [13] Li, S. and Cui, Y. (2012). Gene-centric gene-gene interaction: A model-based kernel machine method. *Annals of Applied Statistics* **6(3)**, 1134–1161.
- [14] Lin, X. (1997). Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika* **84(2)**, 309–326.
- [15] Liu, D. J. and Leal, S. M. (2010). A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet* **6(10)**, e1001156.
- [16] Liu, D. and Leal, S.M. (2012). A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data. *European Journal of Human Genetics* **20**, 449–456.
- [17] Maity, A. and Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics* **67(4)**, 1271–1284.
- [18] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorf, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttman, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461(7265)**, 747–753.
- [19] Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* **1**, 138–147.
- [20] Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., and Cohen, J. C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics* **39**, 513–516.
- [21] Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *The Journal of Clinical Investigation* **119**, 70–79.
- [22] Rotar, V. I. (1974). Some limit theorems for polynomials of second degree. *Theory of Probability and Its Applications* **18(3)**, 499–507.
- [23] Tzeng, J. Y., Zhang, D., Chang, S. M., Thomas, D. C., and Davidian, M. (2009). Gene-Trait Similarity Regression for Multimarker-based Association Analysis. *Biometrics* **65(3)**, 822–832.

- [24] Tzeng, J. Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., Worrall, B. B., Hsu, F. C., Thomas, D. C., and Sullivan, P. F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics* **89(2)**, 277–288.
- [25] Victor, R. G., Haley, R. W., Willett, D. L., Peshock, R. M., Vaeth, P. C., Leonard, D., Basit, M., Cooper, R. S., Iannacchione, V. G., Visscher, W. A., Staab, J. M., Hobbs, H. H. and Dallas Heart Study Investigators (2004). The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American journal of cardiology* **93(12)**, 1473–1480.
- [26] Wei, C., Schaid, D. J., and Lu, Q. (2013). Trees Assembling Mann-Whitney Approach for Detecting Genome-wide Joint Association among Low-Marginal-Effect loci. *Genetic Epidemiology* **37(1)**, 84–91.
- [27] Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics* **86(6)**, 929–942.
- [28] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89(1)**, 82–93.